Ridge Regression

Nipun Batra

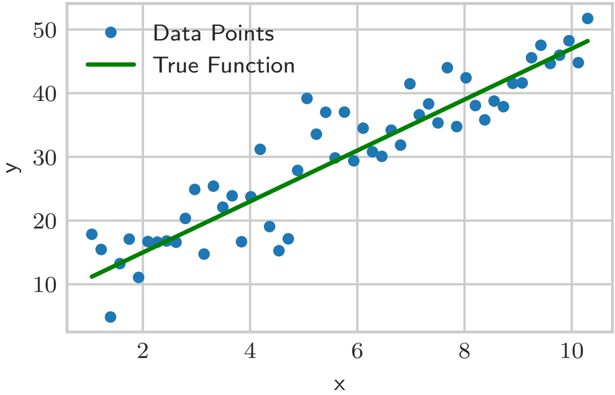
July 23, 2025

IIT Gandhinagar

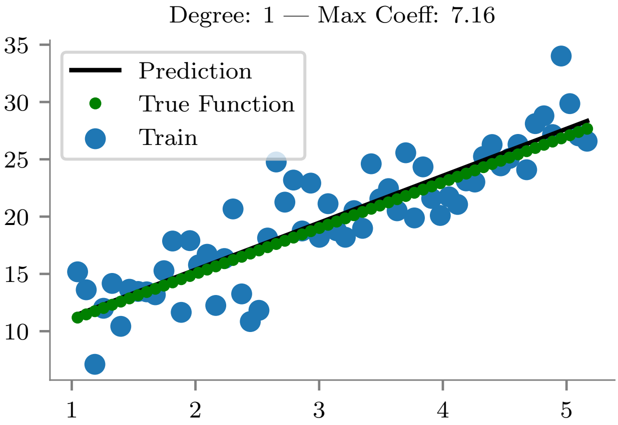
A known measure of overfitting can be the magnitude of the coefficient.

A known measure of overfitting can be the magnitude of the coefficient.

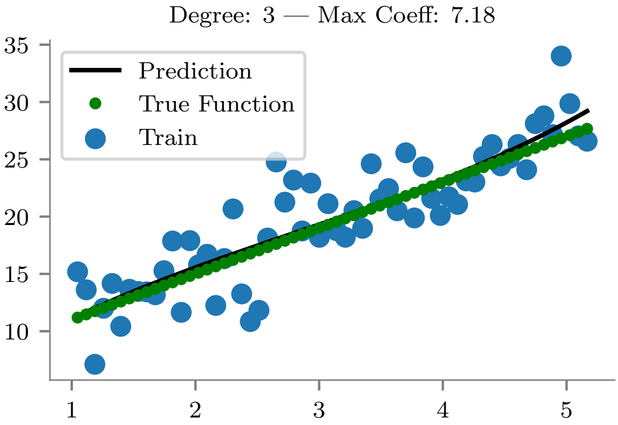
In $f(x) = c_0 + c_1 x + c_2 x^2 + \dots$ it is $\max |c_i|$



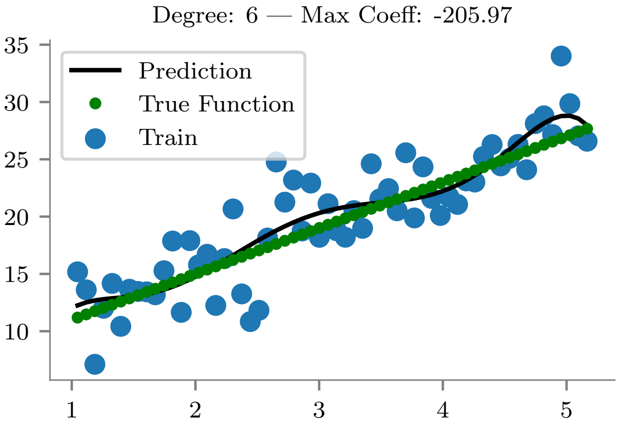
Base Data Set



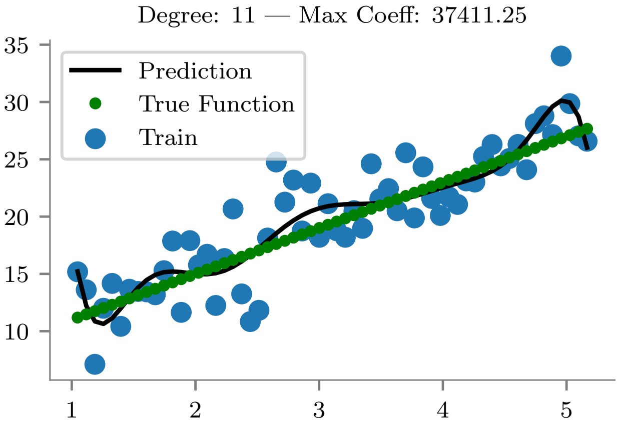
Fit with Degree 1



Fit with Degree 3

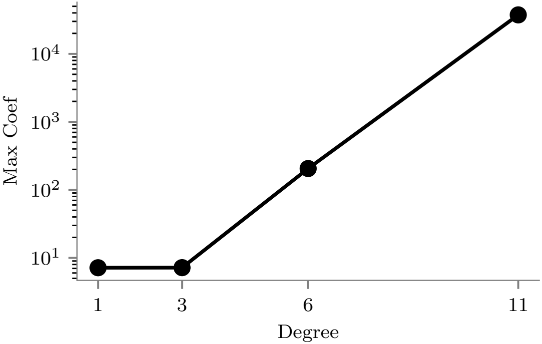


Fit with Degree 6



Fit with Degree 11

In the examples we notice that as the degree increases (as the prediction starts to overfit the base data), the maximum coefficient also increases.



Trend of the coefficients

Introduction

To prevent overfitting we place penalties on large θ_i

To prevent overfitting we place penalties on large θ_i

Objective:

$$\begin{aligned}\text{Minimise } & (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \\ \text{s.t. } & \boldsymbol{\theta}^T \boldsymbol{\theta} \leq S\end{aligned}$$

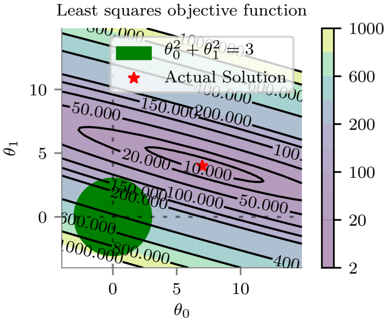
To prevent overfitting we place penalties on large θ_i

Objective:

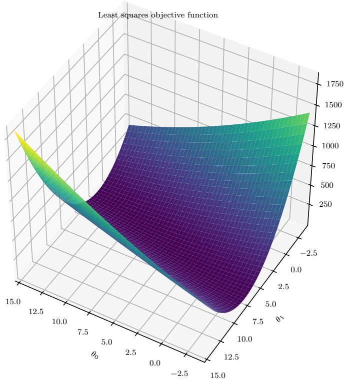
$$\text{Minimise } (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$$
$$\text{s.t. } \boldsymbol{\theta}^T \boldsymbol{\theta} \leq S$$

This is equivalent to

$$\text{Minimise } (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \delta^2 \boldsymbol{\theta}^T \boldsymbol{\theta}$$



(a) Contour Plot



(b) Surface Plot

Visualization of the Example

KKT Conditions

To implement this we use KKT Conditions

To implement this we use KKT Conditions

$$\text{Minimise } \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right)^T \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right)$$
$$\text{s.t. } \boldsymbol{\theta}^T \boldsymbol{\theta} \leq S$$
$$L\left(\boldsymbol{\theta}, \mu\right) = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right)^T \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right) + \mu \left(\boldsymbol{\theta}^T \boldsymbol{\theta} - S\right)$$

where, $\mu \geq 0$ (and $\mu = \delta^2$)

KKT Conditions

To implement this we use KKT Conditions

$$\text{Minimise } (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$$

$$\text{s.t. } \boldsymbol{\theta}^T \boldsymbol{\theta} \leq S$$

$$L(\boldsymbol{\theta}, \mu) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \mu \left(\boldsymbol{\theta}^T \boldsymbol{\theta} - S\right)$$

where, $\mu \geq 0$ (and $\mu = \delta^2$)

KKT Conditions

To implement this we use KKT Conditions

$$\text{Minimise } \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right)^T \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right)$$

$$\text{s.t. } \boldsymbol{\theta}^T \boldsymbol{\theta} \leq S$$

$$L(\boldsymbol{\theta}, \mu) = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right)^T \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right) + \mu \left(\boldsymbol{\theta}^T \boldsymbol{\theta} - S\right)$$

where, $\mu \geq 0$ (and $\mu = \delta^2$)

If $\mu = 0$

There is no regularization

No effect on constraint

KKT Conditions

To implement this we use KKT Conditions

$$\text{Minimise } (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$$
$$\text{s.t. } \boldsymbol{\theta}^T \boldsymbol{\theta} \leq S$$
$$L(\boldsymbol{\theta}, \mu) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \mu \left(\boldsymbol{\theta}^T \boldsymbol{\theta} - S\right)$$

where, $\mu \geq 0$ (and $\mu = \delta^2$)

If $\mu = 0$
There is no regularization
No effect on constraint

KKT Conditions

To implement this we use KKT Conditions

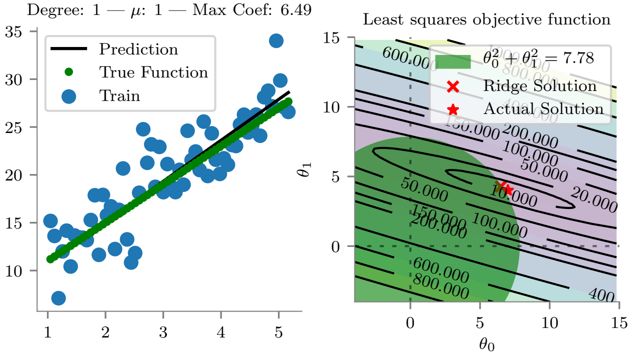
$$\text{Minimise } (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$$
$$\text{s.t. } \boldsymbol{\theta}^T \boldsymbol{\theta} \leq S$$
$$L(\boldsymbol{\theta}, \mu) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \mu \left(\boldsymbol{\theta}^T \boldsymbol{\theta} - S\right)$$

where, $\mu \geq 0$ (and $\mu = \delta^2$)

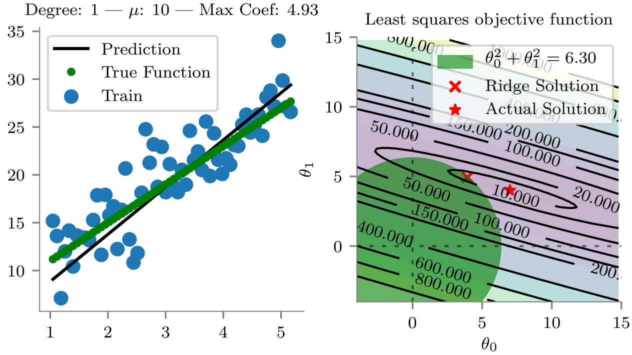
If $\mu = 0$
There is no regularization
No effect on constraint

If $\mu \neq 0$
$\implies \boldsymbol{\theta}^T \boldsymbol{\theta} - S = 0$

Effect of μ

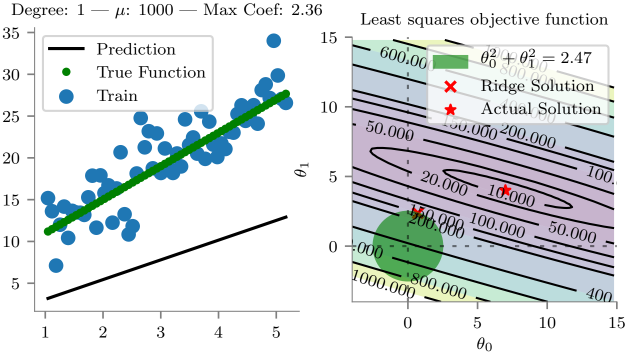


$\mu = 1$


Degree: 1 — μ: 10 — Max Coef: 4.93
Prediction
True Function
Train
Least squares objective function
θ0² + θ1² = 6.30
Ridge Solution
Actual Solution
θ0
θ1

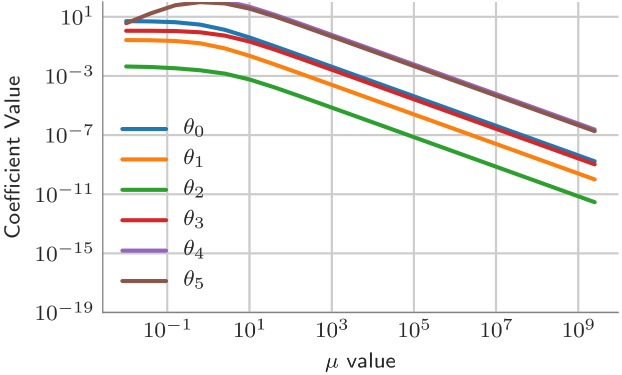
$\mu = 10$

Effect of μ



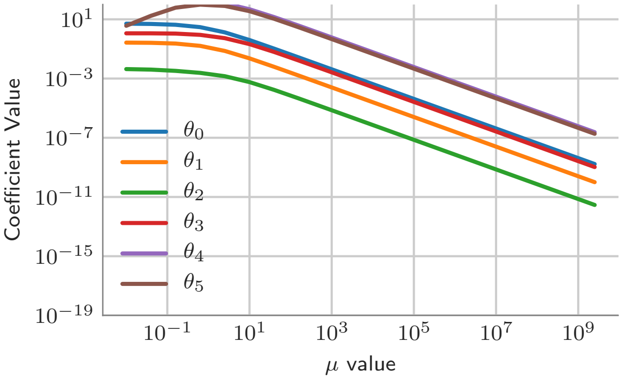
$\mu = 1000$

Effect of μ - Regularization of Parameters



Comparing the magnitudes of the coefficients with varying μ (on the *Real Estate Data Set*)

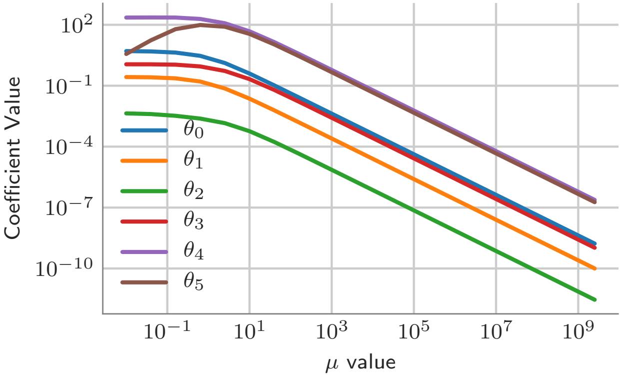
Effect of μ - Regularization of Parameters



Comparing the magnitudes of the coefficients with varying μ (on the *Real Estate Data Set*)

Are θ_i all zero for high μ?

Effect of μ - Regularization of Parameters



Comparing the magnitudes of the coefficients with varying μ
(on the *Real Estate Data Set*)

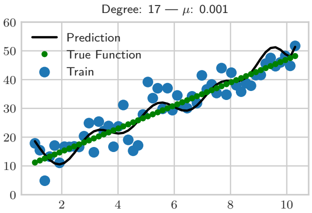
Ridge Objective:

$$\min_{\boldsymbol{\theta}} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right)^T \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right) + \mu\boldsymbol{\theta}^T\boldsymbol{\theta}$$

$$\begin{aligned}
\frac{\partial L\left(\boldsymbol{\theta}, \mu\right)}{\partial \boldsymbol{\theta}} &= 0 \\
\frac{\partial}{\partial \boldsymbol{\theta}} \left\{\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}\right\} + \frac{\partial}{\partial \boldsymbol{\theta}}\mu\boldsymbol{\theta}^T\boldsymbol{\theta} &= 0 \\
\implies -\mathbf{X}^T\mathbf{y} + \left(\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}\right)\boldsymbol{\theta} &= 0 \\
\implies \boldsymbol{\theta}^* &= \left(\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}$$

Bias/Variance

Bias/Variance

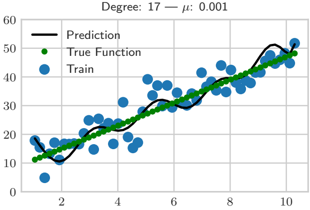


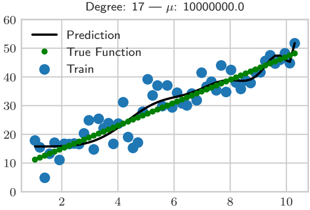
Fit High Order Polynomial

$\implies$ high variance

$\implies$ $\mu \to 0$

Bias/Variance





Fit High Order Polynomial
$\implies$ high variance
$\implies$ $\mu \to 0$

Fit High Order Polynomial
$\implies$ low variance
$\implies$ $\mu \to \infty$

Q.) Solve Regularized ($\mu = 2$) and Unregularized.

Example: Unregularised

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y})$$

Example: Unregularised

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y})$$

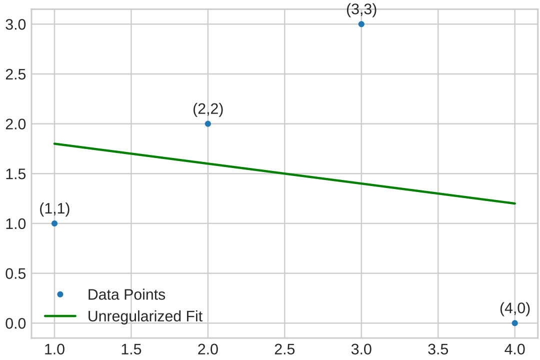
$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{20}\begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

Example: Unregularised

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y})$$
$$\begin{bmatrix}\theta_0\\ \theta_1\end{bmatrix} = \begin{bmatrix}2\\ (-1/5)\end{bmatrix}$$



Example: Regularised

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X} + \mu\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{y})$$

Example: Regularised

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X} + \mu\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{y})$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{X} + \mu\mathbf{I} = \begin{bmatrix} 6 & 10 \\ 10 & 32 \end{bmatrix}$$

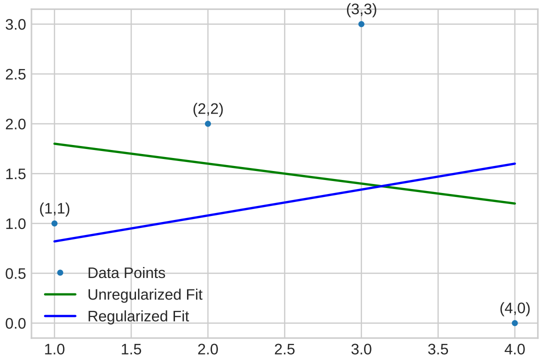
$$(\mathbf{X}^T\mathbf{X} + \mu\mathbf{I})^{-1} = \frac{1}{92}\begin{bmatrix} 32 & -10 \\ -10 & 6 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

Example: Regularised

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X} + \mu\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{y})$$

$$\begin{bmatrix}\theta_0 \\ \theta_1\end{bmatrix} = \begin{bmatrix}0.56 \\ 0.26\end{bmatrix}$$



Multi-collinearity

$(\mathbf{X}^T\mathbf{X})^{-1}$ is not computable when $|\mathbf{X}^T\mathbf{X}| = 0$.

This was a drawback of using linear regression

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{bmatrix}$$

The matrix $\mathbf{X}$ is not full rank.

But with ridge regression, the matrix to be inverted is $\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}$ and not $\mathbf{X}^T\mathbf{X}$.

$$\mathbf{X}^T\mathbf{X} + \mu\mathbf{I} = \begin{bmatrix} 3+\mu & 6 & 12 \\ 6 & 14+\mu & 28 \\ 12 & 28 & 56+\mu \end{bmatrix}$$

The matrix $\mathbf{X}^T\mathbf{X}$ would be full rank for $\mu > 0$.

But with ridge regression, the matrix to be inverted is $\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}$ and not $\mathbf{X}^T\mathbf{X}$.

$$\mathbf{X}^T\mathbf{X} + \mu\mathbf{I} = \begin{bmatrix} 3+\mu & 6 & 12 \\ 6 & 14+\mu & 28 \\ 12 & 28 & 56+\mu \end{bmatrix}$$

The matrix $\mathbf{X}^T\mathbf{X}$ would be full rank for $\mu > 0$.

Another interpretation of “regularisation”

For ridge with no penalty on θ_0

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}^*\right)^{-1}\mathbf{X}^T\mathbf{y}$$

where,

$$\mathbf{I}^* = \begin{bmatrix} \color{red}{0} & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

Ridge Solution using Gradient Descent

- $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha \frac{\partial}{\partial \boldsymbol{\theta}} ((\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \mu \boldsymbol{\theta}^\top \boldsymbol{\theta})$

Ridge Solution using Gradient Descent

- $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha \frac{\partial}{\partial \boldsymbol{\theta}}((\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \mu \boldsymbol{\theta}^\top \boldsymbol{\theta})$
- $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha(-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta} + 2\mu \mathbf{I}\boldsymbol{\theta})$

Ridge Solution using Gradient Descent

- $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha \frac{\partial}{\partial \boldsymbol{\theta}}((\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^{\top} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \mu \boldsymbol{\theta}^{\top} \boldsymbol{\theta})$
- $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha(-2\mathbf{X}^{\top}\mathbf{y} + 2\mathbf{X}^{\top}\mathbf{X}\boldsymbol{\theta} + 2\mu\mathbf{I}\boldsymbol{\theta})$
- $\boldsymbol{\theta} = (1 - 2\alpha\mu\mathbf{I})\boldsymbol{\theta} - \alpha(-2\mathbf{X}^{\top}\mathbf{y} + 2\mathbf{X}^{\top}\mathbf{X}\boldsymbol{\theta})$

Ridge Solution using Gradient Descent

- $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha\frac{\partial}{\partial\boldsymbol{\theta}}((\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \mu\boldsymbol{\theta}^\top\boldsymbol{\theta})$
- $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha(-2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta} + 2\mu\mathbf{I}\boldsymbol{\theta})$
- $\boldsymbol{\theta} = (1 - 2\alpha\mu\mathbf{I})\boldsymbol{\theta} - \alpha(-2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta})$
- $\boldsymbol{\theta} = \underbrace{(1 - 2\alpha\mu\mathbf{I})\boldsymbol{\theta}}_{\text{Shrinking } \boldsymbol{\theta}} - \alpha(-2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta})$

Ridge Solution using Gradient Descent

- $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha \frac{\partial}{\partial \boldsymbol{\theta}}((\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \mu \boldsymbol{\theta}^\top \boldsymbol{\theta})$
- $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha(-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta} + 2\mu \mathbf{I}\boldsymbol{\theta})$
- $\boldsymbol{\theta} = (1 - 2\alpha\mu\mathbf{I})\boldsymbol{\theta} - \alpha(-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta})$
- $\boldsymbol{\theta} = \underbrace{(1 - 2\alpha\mu\mathbf{I})\boldsymbol{\theta}}_{\text{Shrinking } \boldsymbol{\theta}} - \alpha(-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta})$

Ridge Solution using Gradient Descent

- $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha \frac{\partial}{\partial \boldsymbol{\theta}}((\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \mu \boldsymbol{\theta}^\top \boldsymbol{\theta})$
- $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha(-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta} + 2\mu \mathbf{I}\boldsymbol{\theta})$
- $\boldsymbol{\theta} = (1 - 2\alpha\mu\mathbf{I})\boldsymbol{\theta} - \alpha(-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta})$
- $\boldsymbol{\theta} = \underbrace{(1 - 2\alpha\mu\mathbf{I})\boldsymbol{\theta}}_{\text{Shrinking } \boldsymbol{\theta}} - \alpha(-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta})$
- Contrast with update equation for unregularised regression:

- $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha \frac{\partial}{\partial \boldsymbol{\theta}}((\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \mu \boldsymbol{\theta}^\top \boldsymbol{\theta})$
- $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha(-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta} + 2\mu \mathbf{I}\boldsymbol{\theta})$
- $\boldsymbol{\theta} = (1 - 2\alpha\mu\mathbf{I})\boldsymbol{\theta} - \alpha(-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta})$
- $\boldsymbol{\theta} = \underbrace{(1 - 2\alpha\mu\mathbf{I})\boldsymbol{\theta}}_{\text{Shrinking } \boldsymbol{\theta}} - \alpha(-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta})$
- Contrast with update equation for unregularised regression:
- $\boldsymbol{\theta} = \underbrace{\boldsymbol{\theta}}_{\text{No Shrinking } \boldsymbol{\theta}} - \alpha(-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta})$